Journal of the Korean Statistical Society (2006), 35: 4, pp 409–417

THRESHOLD MODELING FOR BIFURCATING AUTOREGRESSION AND LARGE SAMPLE ESTIMATION†

S. Y. HWANG[1] AND SUNGDUCK LEE[2]

ABSTRACT

This article is concerned with threshold modeling of the bifurcating autoregressive model (BAR) originally suggested by Cowan and Staudte (1986) for tree structured data of cell lineage study where each individual (X_t) gives rise to two off-spring (X_{2t}, X_{2t+1}) in the next generation. The triplet (X_t, X_{2t}, X_{2t+1}) refers to mother-daughter relationship. In this paper we propose a threshold model incorporating the difference of "fertility" of the mother for the first and second off-springs, and thereby extending BAR to threshold-BAR (TBAR, for short). We derive a sufficient condition of stationarity for the suggested TBAR model. Also various inferential methods such as least squares (LS), maximum likelihood (ML) and quasi-likelihood (QL) methods are discussed and relevant limiting distributions are obtained.

AMS 2000 subject classifications. Primary 62M10; Secondary .
Keywords. Bifurcating model, fertility, least squares, quasi-likelihood, threshold model.

1. INTRODUCTION

The bifurcating autoregressive model (BAR) was originally suggested by Cowan and Staudte (1986) for tree structured data of cell lineage study. Beginning with the starting value X_1, BAR process $\{X_t,\ t = 1, 2, \ldots\}$ is defined by the following equation

$$X_t = \phi X_{[t/2]} + \epsilon_t, \quad |\phi| < 1, \quad t \geq 2, \tag{1.1}$$

Received May 2006; accepted August 2006.
†This work was supported by grant from KRF(2004-015-C00071).
[1]Corresponding author. Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea (e-mail: shwang@sookmyung.ac.kr)
[2]Department of Information & Statistics, Basic Science Research Institute, Chungbuk National University, Cheongju 361-763, Korea

where $[t]$ denotes the largest integer valued function. For illustration, one can write recursively

$$X_2 = \phi X_1 + \epsilon_2,$$
$$X_3 = \phi X_1 + \epsilon_3,$$
$$X_4 = \phi X_2 + \epsilon_4, \ etc.$$

Here $(\epsilon_{2t}, \epsilon_{2t+1})$ is *iid* bivariate random vector with common mean zero, common variance $\sigma_\epsilon^2 > 0$ and $corr(\epsilon_{2t}, \epsilon_{2t+1}) = \rho$. For statistical analysis of BAR model, refer to Huggins and Staudte (1994), Basawa and Zhou (2004), Zhou and Basawa (2005) and Hwang and Basawa (2006) among others.

In this short paper, we extend BAR model by introducing "threshold effect". The proposed model is formulated by

$$X_t = \phi_t X_{[t/2]} + \epsilon_t, \tag{1.2}$$

where ϕ_t represents binary constant defined by $\phi_t = \phi_1$, for even t and $\phi_t = \phi_2$ when t is odd numbers. Observe that

$$X_2 = \phi_1 X_1 + \epsilon_2,$$
$$X_3 = \phi_2 X_1 + \epsilon_3,$$
$$X_4 = \phi_1 X_2 + \epsilon_4,$$
$$X_5 = \phi_2 X_2 + \epsilon_5, \ etc.$$

The triplet (X_t, X_{2t}, X_{2t+1}) is referred to as mother-daughter relationship. It is noted that each individual (X_t) produces two off-spring (X_{2t}, X_{2t+1}) in the next generation in such a manner that X_{2t} and X_{2t+1} are related to distinctive autoregressive coefficients ϕ_1 and ϕ_2 respectively. Accordingly, in the proposed threshold-BAR model, ϕ_1 and ϕ_2 can be viewed as "fertility" of the mother and the first and second off-springs are associated with "fertility" ϕ_1 and ϕ_2, respectively and therby extending BAR to threshold-BAR (TBAR, hereafter) model. An illustrative Figure 1.1 will help understand the threshold-binary structure of the three structured data suitable for TBAR model.

This paper derives a stationarity condition for TBAR model. Also various inferential problems such as least squares (LS), maximum likelihood (ML) and quasi-likelihood (QL) methods are discussed and relevant limiting distributions are derived. Notice that ML requires specification of the error distributions while LS and QL methods can be applied without distributional assumptions.

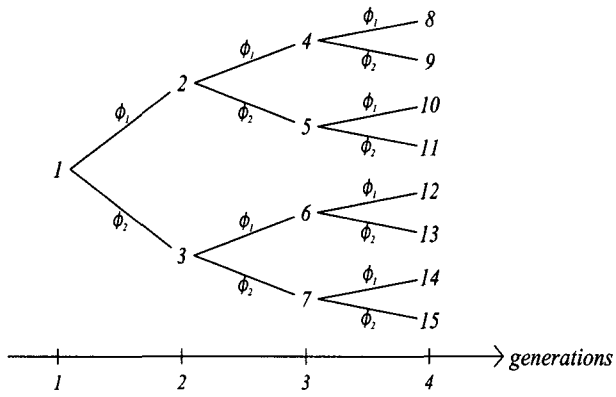

FIGURE 1.1 *Threshold tree-structured data.*

2. The Model and Stationarity Condition

Consider the following TBAR process specified by

$$X_t = \phi_t X_{[t/2]} + \epsilon_t,$$

where the coefficient ϕ_t is binary, taking values ϕ_1 and ϕ_2 according to the even numbers and odd numbers of t, respectively. Here $(\epsilon_{2t}, \epsilon_{2t+1})$ is *iid* bivariate random vector with common mean zero, common variance $\sigma_\epsilon^2 > 0$ and $corr(\epsilon_{2t}, \epsilon_{2t+1}) = \rho$. Consider the ancestral path $a(t)$ of (X_{2t}, X_{2t+1}), given by

$$a(t) = \{X_{[t/2^j]},\ j = 0, 1, 2, \ldots\}.$$

Define the following augmented ancestral path $A(t)$ of (X_{2t}, X_{2t+1}) including sisters of $X_{[t/2^j]}$, $j = 0, 1, 2, \ldots$ for each generation. Specifically,

$$A(t) = \{X_{[t/2^j]}, X_{[t/2^j]+(-1)^{[t/2^j]}},\ j = 0, 1, 2, \ldots\}. \tag{2.1}$$

For illustration, take $t = 7$ and note that augmented ancestral path $A(7)$ of (X_{14}, X_{15}) is given by

$$A(7) = \{X_7, X_6, X_3, X_2, X_1\}.$$

We first establish stationarity and geometric ergodicity of the model.

(C1) The common bivariate distribution of $(\epsilon_{2t}, \epsilon_{2t+1})$ is absolutely continuous with respect to Lebesgue measure in R^2.

THEOREM 2.1. *Assume (C1). If* $max(|\phi_1|, |\phi_2|) < 1$, *then the proposed TBAR model permits a unique stationary distribution and the process is geometrically ergodic along the augmented ancestral path* $A(t)$.

REMARK. When $\phi_1 = \phi_2 = \phi$, TBAR model reduces to standard BAR in (1.1) and the stationarity condition $\max(|\phi_1|, |\phi_2|) < 1$ reduces to $|\phi| < 1$ which is the stationarity condition for standard BAR in (1.1). Bivariate normal distribution for $(\epsilon_{2t}, \epsilon_{2t+1})$ clearly satisfies (C1). It is conjectured that Theorem 2.1 continues to be valid even for non-continuous type distributions for $(\epsilon_{2t}, \epsilon_{2t+1})$ such as bivariate Poisson.

PROOF. First fix t and for given augmented ancestral path $A(t)$, consider the bivariate time series

$$Y_j = (X_{[t/2^j]}, X_{[t/2^j]+(-1)^{[t/2^j]}})^T, \ j = 0, 1, 2, \ldots$$

Notice that Y_j can be written in terms of a bivariate Markovian AR(1) process as

$$Y_j = \Phi Y_{j-1} + (\epsilon_{[t/2^j]}, \epsilon_{[t/2^j]+(-1)^{[t/2^j]}})^T, \ j = 0, 1, 2, \ldots, \tag{2.2}$$

where for odd $[t/2^j]$

$$\Phi = \begin{pmatrix} \phi_1 & 0 \\ \phi_2 & 0 \end{pmatrix}$$

and for even $[t/2^j]$

$$\Phi = \begin{pmatrix} \phi_2 & 0 \\ \phi_1 & 0 \end{pmatrix}.$$

Consequently, the proof follows from standard arguments for Markovian vector time series such as those in Feigin and Tweedie (1985). The condition (C1) and $\max(|\phi_1|, |\phi_2|) < 1$ give a set of sufficient conditions for the condition in Theorem 1 of Feigin and Tweedie (1985). Also, Irreducibility and Feller-Chain follows from (C1) since transition function of the TBAR model is continuous, completing the proof. □

It will be assumed throughout that result in Theorem 2.1 holds.

3. Estimation of Parameters

For the estimation of parameters in the context of time series models, the likelihood function is typically unknown or it is too complicated for practical purposes. In such cases it is a common practice to employ least squares (LS) method obtained by minimizing the error sum of squares. Alternatively, one can specify the "objective function" only using the first and second order moments. Quasi-likelihood (QL) method often provides a systematic and unified approach for inference for partially specified model rather than likelihood-specified model. See Godambe (1985). It is noted that QL approach is referred to under various names such as estimating function (EE), martingale estimating equation (MEE) and generalized methods of moments (GMM). In this section we first discuss LS and QL methods without specification of error distribution, *i.e.*, without knowing the likelihood of the data. ML estimation will also be discussed later. Denote the number of mother-daughter triplet (X_t, X_{2t}, X_{2t+1}) by n so that we have total number of observation $N = 2n + 1$.

3.1. LS estimation

Note that one can write error sum of squares

$$Q = \sum (X_{2t} - \phi_1 X_t)^2 + \sum (X_{2t+1} - \phi_2 X_t)^2,$$

where and in what follows the summation runs from $t = 1$ and $t = n$. The least squares estimator $\widehat{\phi_{LS}} = (\widehat{\phi_{1LS}}, \widehat{\phi_{2LS}})^T$ of $\phi = (\phi_1, \phi_2)^T$ is seen to be

$$\widehat{\phi_{1LS}} = \frac{\sum X_{2t} X_t}{\sum X_t^2}$$

and

$$\widehat{\phi_{2LS}} = \frac{\sum X_{2t+1} X_t}{\sum X_t^2}.$$

Define two matrices A and B given by

$$A = \operatorname{plim}\left[\frac{n^{-1}\partial^2 Q}{\partial \phi^2}\right]$$

and $B =$ asymptotic variance-covariance matrix of $n^{-1/2}\partial Q/\partial \phi$.

(C2) Finite fourth order moment, *i.e.*, $EX_t^4 < \infty$.

THEOREM 3.1. *Under (C1) and (C2), we have as n goes to infinity*

$$\sqrt{n}(\widehat{\phi_{LS}} - \phi) \quad \xrightarrow{d} \quad N(0, A^{-1}BA^{-1}),$$

where A and B are defined above.

PROOF. Following the lines as in Klimko and Nelson (1978), one can deduce

$$\sqrt{n}(\widehat{\phi_{LS}} - \phi) = -\left[\frac{n^{-1}\partial^2 Q}{\partial\phi^2}\right]^{-1}\left[\frac{n^{-1/2}\partial Q}{\partial\phi}\right] + o_p(1). \tag{3.1}$$

Martingale central limit theorem (see for instance Corollary 3.1 of Hall and Heyde (1980), p. 58) gives, due to (C2),

$$\frac{n^{-1/2}\partial Q}{\partial\phi} \quad \xrightarrow{d} \quad N(0, B) \tag{3.2}$$

and via the ergodic theorem one can obtain

$$\frac{n^{-1}\partial^2 Q}{\partial\phi^2} \quad \xrightarrow{p} \quad A. \tag{3.3}$$

Combining (3.1) to (3.3) entails the theorem. □

3.2. QL estimation

Direct calculation yields for our TBAR model

$$E(X_{2t}|X_t) = \phi_1 X_t, \quad E(X_{2t+1}|X_t) = \phi_2 X_t$$

$$\mathrm{Var}(X_{2t}|X_t) = \mathrm{Var}(X_{2t+1}|X_t) = \sigma_\epsilon^2$$

$$\mathrm{Cov}(X_{2t}, X_{2t+1}|X_t) = \rho\sigma_\epsilon^2.$$

Also define

$$m_t = (X_{2t} - \phi_1 X_t, X_{2t+1} - \phi_2 X_t)^T$$

and

$$V = \sigma_\epsilon^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

The quasi-likelihood estimator $\widehat{\phi_{QL}} = (\widehat{\phi_{1QL}}, \widehat{\phi_{2QL}})^T$ of $\phi = (\phi_1, \phi_2)^T$ is then obtained by solving the following quasi-likelihood equation:

$$\sum \partial m_t^T V^{-1} \partial m_t = 0, \tag{3.4}$$

where ∂m_t denotes derivative matrix of m_t with respect to ϕ. See, for instance, Basawa and Zhou (2004). It will be assumed that

$$C = E[\partial m_t^T V^{-1} \partial m_t] < \infty. \tag{3.5}$$

THEOREM 3.2. *Under (C1) and (C2), we have as n goes to infinity*

$$\sqrt{n}(\widehat{\phi_{QL}} - \phi) \quad \xrightarrow{d} \quad N(0, C^{-1})$$

where C is defined in (3.5).

PROOF. Following the lines as in Godambe (1985) and Basawa and Zhou (2004), it can be shown that

$$\sqrt{n}(\widehat{\phi_{QL}} - \widehat{\phi}) = -[n^{-1} \sum \partial^2 m_t^T V^{-1} \partial^2 m_t]^{-1} [n^{-1/2} \sum \partial m_t^T V^{-1} \partial m_t] + o_p(1) \tag{3.6}$$

Here $\partial^2 m_t$ denotes second order derivative matrix of m_t, with respect to ϕ. The first factor in the RHS of (3.6) converges in probability to D^{-1} where

$$D = -E(\partial^2 m_t^T V^{-1} \partial^2 m_t). \tag{3.7}$$

Also, martingale central limit theorem tells us that the second factor in the RHS of (3.6) converges in distribution to $N(0, C)$ with C defined in(3.5). Moreover, argument similar to that in Eq. (3.12) of Hwang and Basawa (2003) reveals $D = C$, which in turn implies the theorem using (3.6). □

3.3. ML estimation

The likelihood function $L_n(\phi)$ is obtained by

$$L_n(\phi) = \Pi f(X_{2t}, X_{2t+1} | X_t), \tag{3.8}$$

where $f(X_{2t}, X_{2t+1}|X_t)$ is the bivariate density of (X_{2t}, X_{2t+1}) conditionally on X_t. It will be assumed that $L_n(\phi)$ is completely known. The score vector and Hessian matrix are defined as the first order derivative vector and the minus the second order matrix of the log-likelihood, respectively. Specifically,

$$S_n(\phi) = \frac{\partial \log L_n(\phi)}{\partial \phi}$$

and

$$H_n(\phi) = -\frac{\partial S_n(\phi)}{\partial \phi}.$$

The ML estimator $\widehat{\phi_{ML}}$ is typically obtained by solving the likelihood equation $S_n(\phi) = 0$. Sometimes, the explicit solution may not be available. Even in this case, one can continue to define $\widehat{\phi_{ML}}$ via one-step iteration given by

$$\widehat{\phi_{ML}} = \phi_0 + H_n^{-1}(\phi_0)S_n(\phi_0), \tag{3.9}$$

where ϕ_0 is a preliminary $\sqrt{n}-$consistent estimator of ϕ. One may take $\widehat{\phi_{LS}}$ and $\widehat{\phi_{QL}}$ in place of ϕ_0.

(C3) There exist nonsingular matrix $F(\phi)$ such that

$$F(\phi) = \text{plim}[n^{-1}H_n(\phi)]. \tag{3.10}$$

THEOREM 3.3. *Under (C1) and (C3), we have*

$$\sqrt{n}(\widehat{\phi_{ML}} - \phi) \quad \xrightarrow{d} \quad N(0, F^{-1}(\phi)).$$

PROOF. By the Taylor's expansion of $S_n(\phi)$ about $\phi = \widehat{\phi_{ML}}$ one can easily see that

$$n^{-1/2}S_n(\phi) = n^{-1}H_n(\phi)\sqrt{n}(\widehat{\phi_{ML}} - \phi) + o_p(1). \tag{3.11}$$

One can also show via the martingale central limit theorem that $n^{-1/2}S_n(\phi)$ converges in distribution to $N(0, F(\phi))$. Thus, theorem follows from (3.10) and (3.11). Above arguments hold only for the case that $S_n(\widehat{\phi_{ML}}) = 0$. When $\widehat{\phi_{ML}}$ is near zero of the likelihood equation, it can be shown, omitting details, via quadratic mean differentiability arguments (Hwang and Basawa, 1993) that the result in the theorem continues to be valid. □

ACKNOWLEDGEMENT

We thank the two referees for careful reading of the paper. This work was supported by a grant from KRF(2004-015-C00071).

REFERENCES

BASAWA, I. V. AND ZHOU, J. (2004). "Non-Gaussian bifurcating models and quasi-likelihood estimation", *Journal of Applied Probability*, **41A**, 55–64.

COWAN, R. AND STAUDTE, R. G. (1986). "The bifurcating autoregression model in cell lineage studies", *Biometrics*, **42**, 769–783.

FEIGIN, P. D. AND TWEEDIE, R. L. (1985). "Random coefficient autoregressive processes: A Markov chain analysis of stationarity and finiteness of moments", *Journal of Time Series Analysis*, **6**, 1–14.

GODAMBE, V. P. (1985). "The foundations of finite sample estimation in stochastic processes", *Biometrika*, **72**, 419–428.

HALL, P. G. AND HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*, Academic Press, New York.

HUGGINS, R. M. AND STAUDTE, R. G. (1994). "Variance components models for dependent cell populations", *Journal of the American Statistical Association*, **89**, 19–29.

HWANG, S. Y. AND BASAWA, I. V. (1993). "Asymptotic optimal inference for a class of nonlinear time series models", *Stochastic Processes and Their Applications*, **46**, 91–113.

HWANG, S. Y. AND BASAWA, I. V. (2003). "Estimation for nonlinear autoregressive models generated by beta-ARCH processes", *Sankhyā*, **65**, 744–762

HWANG, S. Y. AND BASAWA, I. V. (2006). "Local asymptotic normality for bifurcating autoregressive processes and related asymptotic inference", Technical Report #2006-04, University of Georgia.

KLIMKO, L. A. AND NELSON, P. I. (1978). "On conditional least squares estimation for stochastic processes", *The Annals of Statistics*, **6**, 629–642.

ZHOU, J. AND BASAWA, I. V. (2005). "Maximum likelihood estimation for first order bifurcating autoregressive process with exponential errors", *Journal of Time Series Analysis*, **26**, 825–842.